(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2017 年 8 月 10 日 (10.08.2017)

WIPO | PCT

(10) 国际公布号
WO 2017/133233 A1

(51) 国际专利分类号:
H04L 29/08 (2006.01)

(21) 国际申请号: PCT/CN2016/097244

(22) 国际申请日: 2016 年 8 月 29 日 (29.08.2016)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201610082068.8 2016 年 2 月 5 日 (05.02.2016) CN

(71) 申请人: **华为技术有限公司 (HUAWEI TECHNOLOGIES CO., LTD.)** [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(72) 发明人: **刘存伟 (LIU, Cunwei)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 **吴国军 (WU, Guojun)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 **黄西华 (HUANG, Xihua)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 **金雪锋 (JIN, Xuefeng)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(74) 代理人: **北京三高永信知识产权代理有限责任公司 (BEIJING SAN GAO YONG XIN INTELLECTUAL PROPERTY AGENCY CO., LTD.)**; 中国北京市海淀区学院路蓟门里和景园 A 座 1 单元 102 室, Beijing 100088 (CN)。

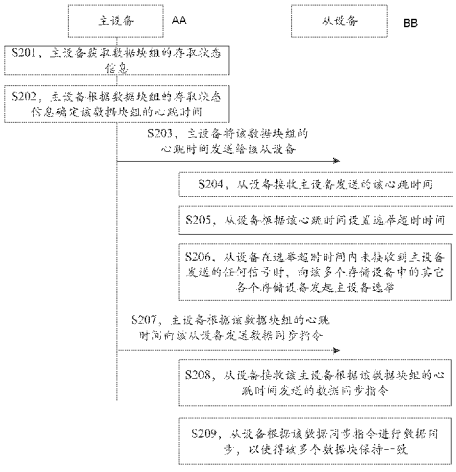
(81) 指定国 (除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明，要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH,

[见续页]

(54) **Title**: HEARTBEAT-BASED DATA SYNCHRONIZATION DEVICE, METHOD, AND DISTRIBUTED STORAGE SYSTEM

(54) **发明名称**: 基于心跳的数据同步装置、方法及分布式存储系统



S201 Master apparatus obtains access status information of the data block set
S202 Master apparatus determines according to the access status information of the data block set a heartbeat time of the data block set
S203 Master apparatus transmits, according to the heartbeat time of the data block set and to the slave apparatus, a synchronization command
S204 Slave apparatus receives the heartbeat time transmitted from the master apparatus
S205 Slave apparatus configures according to the heartbeat time an election expiry time
S206 If no signal is received in the election expiry time, slave apparatus initiates, to other storage apparatuses of the plurality of the storage apparatuses, a master apparatus election procedure
S207 Master apparatus transmits, according to the heartbeat time of the data block set and to the slave apparatus, a data synchronization command
S208 Slave apparatus receives the data synchronization command transmitted according to the heartbeat time by the master apparatus
S209 Slave apparatus performs, according to the data synchronization command, data synchronization to enable a plurality of data blocks to remain consistent
AA Master apparatus
BB Slave apparatus

图 2

(57) **Abstract**: The invention relates to the field of the data synchronization technology, and specifically, to a heartbeat-based data synchronization method. The method is used in a distributed storage system storing at least one data block set and comprising a plurality of storage apparatuses. One of the plurality of storage apparatuses is a master apparatus configured to store the data block set, and the remaining apparatus is a slave apparatus configured to store the data block set. The master apparatus executes the method comprising the steps of: obtaining access status information of the data block set, determining according to the access status information of the data block set a heartbeat time of the data block set, and transmitting, according to the heartbeat time of the data block set and to the slave apparatus, a synchronization command used to instruct the slave apparatus to perform data synchronization. The embodiment achieves the goals of reducing a system overhead of the distributed storage system and increasing read and write performance in the storage system.

(57) **摘要**:

[见续页]

WO 2017/133233 A1

CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

本发明公开了一种基于心跳的数据同步方法，属于数据同步技术领域。所述方法用于分布式存储系统中，所述分布式存储系统存储有至少一个数据块组，且所述分布式存储系统存储包括多个存储设备；所述多个存储设备中的一个设备为存储所述数据块组的主设备，其余设备为存储所述数据块组的从设备，所述主设备执行所述方法，所述主设备获取所述数据块组的存取状态信息，根据所述数据块组的存取状态信息确定所述数据块组的心跳时间，根据所述数据块组的心跳时间向所述从设备发送数据同步指令，所述数据同步指令用于指示所述从设备进行数据同步，达到降低分布式存储系统的系统开销，提高存储系统的读写性能的效果。

基于心跳的数据同步装置、方法及分布式存储系统

技术领域

本发明涉及数据同步技术领域，特别涉及一种基于心跳的数据同步装置、方法及分布式存储系统。

背景技术

在数据同步技术领域中，Raft 一致性算法广泛应用在分布式存储系统中。

目前，在基于 Raft 一致性算法的分布式存储系统中的数据分为若干个数据块组，每个数据块组由分别存储在不同存储设备中的多个相同数据块组成，在实现 Raft 一致性算法时，以数据块组为单位进行数据同步。具体的，对于一个数据块组，该数据块组的各个数据块所在的存储设备首先通过选举的方式产生该数据块组的一个主设备，相应的，其它设备为该数据块组的从设备，其中，主设备负责处理与客户端之间的交互。当主设备接收到客户端发送的读写指令时，将读写指令记录为日志，并按照固定的心跳时间向各个从设备发送包含该日志的数据同步指令，由各个从设备根据该数据同步指令进行数据同步；若在心跳时间达到时没有要发送的日志，则主设备需要向各个从设备发送不包含数据同步指令的心跳信号，以确定连接正常。

在实现本发明的过程中，发明人发现现有技术至少存在以下问题：

现有的 Raft 一致性算法中，各个数据块组的心跳时间都是固定值，而各个数据块组的读写频率却并不均衡，为了保证读写频率较高的数据块组中各个数据块之间能够及时同步，该固定值通常设置的较小，从而导致读写频率较低的数据块组所在的各个存储设备之间也需要频繁的收发信号，导致系统开销较大，影响存储系统的读写性能。

发明内容

为了解决现有技术中读写频率较低的数据块组所在的各个存储设备之间也需要频繁的收发的信号，导致系统开销较大，影响存储系统的读写性能的问题，本发明实施例提供了一种装置、方法及分布式存储系统。所述技术方案如下：

分布式存储系统存储有至少一个数据块组，且该分布式存储系统包括多个存储设备；该多个存储设备中的一个设备为存储该数据块组的主设备，其余设备为存储该数据块组的从设备。该分布式存储系统还可以包括协调设备，该协调设备与分布式存储系统中的各个存储设备相连。

第一方面，提供了一种基于心跳的数据同步方法，该方法包括：

主设备获取该数据块组的存取状态信息；该主设备根据该数据块组的存取状态信息确定该数据块组的心跳时间；该主设备根据该数据块组的心跳时间向该从设备发送数据同步指令，该数据同步指令用于指示该从设备进行数据同步。

本发明实施例所示的方案，对于每一个数据块组，根据该数据块组的读写频率等相关信息自适应的确定数据块组的心跳时间，解决了现有的 Raft 一致性算法中，一个数据块组所在的各个存储设备之间需要频繁的收发的信号，导致系统开销较大，影响存储系统的读写性能的问题；达到了降低分布式存储系统的系统开销，提高存储系统的读写性能的效果。

在第一方面的第一种可能实现方式中，数据块组的存取状态信息包括数据块组的读频率和写频率。

结合第一方面或者第一方面的第一种可能实现方式，在第一方面的第二种可能实现方式中，该主设备根据该数据块组的存取状态信息确定该数据块组的心跳时间，包括：该主设备根据预设的第一评分规则对存取状态信息进行评分，获得第一参考分值；该主设备根据该第一参考分值确定该数据块组的心跳时间。提供一种根据包括读频率和写频率在内的至少两个信息确定心跳时间的方式，自适应的确定数据块组的心跳时间。

结合第一方面的第一种可能实现方式，在第一方面的第三种可能实现方式中，该主设备根据该数据块组的存取状态信息确定该数据块组的心跳时间，包括：该主设备根据该第一评分规则对该读频率和该写频率分别进行评分；该主设备将对该读频率和该写频率的评分的和作为第一参考分值；该主设备根据该第一参考分值确定该数据块组的心跳时间。

结合第一方面的第一种可能实现方式，在第一方面的第四种可能实现方式中，该主设备根据该数据块组的存取状态信息确定该数据块组的心跳时间，包括：该主设备获取预先设置的参考时间以及该存取状态信息对应的第一权重；该主设备根据该存取状态信息、该参考时间以及该第一权重计算该数据块组的心跳时间。提供一种根据包括读频率和写频率在内的至少两个信息确定心跳时

间的方式，自适应的确定数据块组的心跳时间。

结合第一方面的第四种可能实现方式，在第一方面的第五种可能实现方式中，该主设备根据该数据块组的存取状态信息确定该数据块组的心跳时间，包括：

通过以下公式计算该心跳时间：

heartbeatTime= Time/(weightR*R+ weightW*W)；weightR+weightW=1；

其中，heartbeatTime 为该心跳时间，Time 为该参考时间，R 为该读频率的数值，W 为该写频率的数值，weightR 为该读频率对应的权重，weightW 为该写频率对应的权重。

本发明实施例所示的方案，主设备根据该数据块组的存取状态信息确定该数据块组的心跳时间，综合考虑了不同存取状态信息对整体存储系统的影响，从实际的数据存取情况出发，更加合理地配置心跳时间，解决了现有的多 Raft 系统中，一个数据块组所在的各个存储设备之间短时内频繁的收发的信号造成的“心跳风暴”，导致系统开销较大，影响存储系统的读写性能的问题。

结合第一方面或者第一方面的第一至五种可能实现方式中的任意一种，在第一方面的第六种可能实现方式中，该方法还包括：该主设备将该数据块组的心跳时间发送给该从设备，使得该从设备根据该心跳时间设置选举超时时间。

本发明实施例所示的方案，主设备将该数据块组的心跳时间发送给该从设备，使得该从设备根据该心跳时间设置选举超时时间，考虑了心跳时间对选举超时时间的影响，选举出最有利于系统整体性能发挥的主设备主导数据的同步，提高了存储系统的整体读写性能。

结合第一方面或者第一方面的第一至六种可能实现方式中的任意一种，在第一方面的第七种可能实现方式中，该分布式存储系统还包含与所述多个存储设备相连接的协调设备，该方法还包括：该主设备将该数据块组的心跳时间发送给该协调设备；该主设备接收该协调设备返回的、修正后的心跳时间；该主设备根据该修正后的心跳时间向该从设备发送数据同步指令。

本发明实施例所示的方案，根据各个数据块组各自的读写频率对各个数据块组的心跳时间进行修正，从整体上对存储系统各个数据块组的心跳时间进行优化，进一步提高存储系统的读写性能。

第二方面，提供了一种基于心跳的数据同步方法，该方法包括：从设备接

收该主设备根据该数据块组的心跳时间发送的数据同步指令；该心跳时间是由该主设备获取该数据块组的存取状态信息，并根据该数据块组的存取状态信息确定；该从设备根据该数据同步指令进行数据同步。

本发明实施例所示的方案，根据该数据块组的读写频率等相关信息自适应的确定数据块组的心跳时间，根据确定的心跳时间进行设备间的数据同步，解决了现有的 Raft 一致性算法中，一个数据块组所在的各个存储设备之间需要频繁的收发的信号，导致系统开销较大，影响存储系统的读写性能的问题；达到了降低分布式存储系统的系统开销，提高存储系统的读写性能的效果。

在第二方面的第一种可能实现方式中，该数据块组的存取状态信息包括该数据块组的读频率和写频率。

结合第二方面的第一种可能实现方式，在第二方面的第二种可能实现方式中，该方法还包括：该从设备接收该主设备发送的该心跳时间；该从设备根据该心跳时间设置选举超时时间；当该从设备在该选举超时时间内未接收到该主设备发送的任何信号时，该从设备向该多个存储设备中的其它各个存储设备发起主设备选举。

本发明实施例所示的方案，从设备根据该心跳时间设置选举超时时间，考虑了心跳时间对选举超时时间的影响，选举出最有利于系统整体性能发挥的主设备主导数据的同步，提高了存储系统的整体读写性能。

结合第二方面的第二种可能实现方式，在第二方面的第三种可能实现方式中，该从设备根据该心跳时间设置选举超时时间，包括：该从设备根据预设的第二评分规则对该存取状态信息进行评分，获得第二参考分值；该从设备根据该第二参考分值确定该数据块组的第一超时系数；该从设备将该第一超时系数与该心跳时间的乘积设置为该选举超时时间。

结合第二方面的第二种可能实现方式，在第二方面的第四种可能实现方式中，该从设备根据该心跳时间设置选举超时时间，包括：该从设备获取预先设置的参考系数以及所述存取状态信息对应的第二权重；该从设备根据该存取状态信息、该参考系数以及该第二权重计算该数据块组的第二超时系数；该从设备将该第二超时系数与该心跳时间的乘积设置为该选举超时时间。

结合第二方面的第四种可能实现方式，在第二方面的第五种可能实现方式中，该从设备根据该心跳时间设置选举超时时间，包括：

通过以下公式设置该选举超时时间：

OverTime = （weightR*R+ weightW*W+Reference）* heartbeatTime;

weightR+weightW=1;

其中， OverTime 为该选举超时时间， heartbeatTime 为该心跳时间，Reference 为该参考系数，R 为该读频率的数值，W 为该写频率的数值，weightR 为该读频率对应的权重，weightW 为该写频率对应的权重。

本发明实施例所示的方案，主设备根据该数据块组的存取状态信息确定该数据块组的心跳时间，根据该心跳时间设置选举超时时间，综合考虑了不同存取状态信息对整体存储系统的影响，从实际的数据存取情况出发，更加合理地配置选举时间，提高了存储系统整体的读写性能。

第三方面，提供了一种基于心跳的数据同步方法，该方法包括：协调设备统计该数据块组的存取状态信息，并接收该主设备发送的、该数据块组的心跳时间；该心跳时间是由该主设备获取该数据块组的存取状态信息，并根据该数据块组的存取状态信息确定；该协调设备根据该数据块组的存取状态信息确定该数据块组的重要性等级；该协调设备根据该数据块组的重要性等级对该心跳时间进行修正；该协调设备将修正后的心跳时间返回给该主设备。

本发明实施例所示的方案，协调设备根据该数据块组的存取状态信息确定该数据块组的重要性等级，并根据该数据块组的重要性等级对该心跳时间进行修正。从整体系统的角度，通过对心跳时间的统筹调整，更加合理地分配资源，提高了存储系统整体的读写性能。

第四方面，提供了一种网络设备，该网络设备包括：处理器、网络接口、存储器以及总线，存储器与网络接口分别通过总线与处理器相连；处理器被配置为执行存储器中存储的指令；处理器通过执行指令来实现上述第一方面或第一方面中任意一种可能的实现方式所提供的基于心跳的数据同步方法。

第五方面，提供了一种网络设备，该网络设备包括：处理器、网络接口、存储器以及总线，存储器与网络接口分别通过总线与处理器相连；处理器被配置为执行存储器中存储的指令；处理器通过执行指令来实现上述第二方面或第二方面中任意一种可能的实现方式所提供的基于心跳的数据同步方法。

第六方面，提供了一种网络设备，该网络设备包括：处理器、网络接口、存储器以及总线，存储器与网络接口分别通过总线与处理器相连；处理器被配置为执行存储器中存储的指令；处理器通过执行指令来实现上述第三方面所提供的基于心跳的数据同步方法。

第七方面，本发明实施例提供了一种基于心跳的数据同步装置，该界面交互装置包括至少一个单元，该至少一个单元用于实现上述第一方面或第一方面中任意一种可能的实现方式所提供的基于心跳的数据同步方法。

第八方面，本发明实施例提供了一种基于心跳的数据同步装置，该界面交互装置包括至少一个单元，该至少一个单元用于实现上述第二方面或第二方面中任意一种可能的实现方式所提供的基于心跳的数据同步方法。

第九方面，本发明实施例提供了一种基于心跳的数据同步装置，该界面交互装置包括至少一个单元，该至少一个单元用于实现上述第三方面所提供的基于心跳的数据同步方法。

上述本发明实施例第四到第九方面所获得的技术效果与第一到第三方面中对应的技术手段获得的技术效果近似，在这里不再赘述。

综上所述，本发明实施例提供的技术方案带来的有益效果是：

通过主设备获取数据块组的存取状态信息；主设备根据数据块组的存取状态信息确定数据块组的心跳时间；主设备根据数据块组的心跳时间向所从设备发送数据同步指令，数据同步指令用于指示从设备进行数据同步，解决了现有的 Raft 一致性算法中，一个数据块组所在的各个存储设备之间需要频繁的收发的信号，导致系统开销较大，影响存储系统的读写性能的问题；达到了降低分布式存储系统的系统开销，提高存储系统的读写性能的效果。

附图说明

为了更清楚地说明本发明实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明

的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图 1A 是本发明一示例性实施例提供的分布式存储系统的结构示意图；

图 1B 是图 1A 所示实施例涉及的一种网络设备的结构示意图；

图 1C 是图 1A 所示实施例涉及的一种应用程序模块的示意图；

图 1D 是图 1A 所示实施例涉及的另一种应用程序模块的示意图；

图 1E 是图 1A 所示实施例涉及的又一种应用程序模块的示意图；

图 2 是本发明一示例性实施例提供的一种基于心跳的数据同步方法的方法流程图；

图 3 是本发明一示例性实施例提供的一种基于心跳的数据同步方法的方法流程图；

图 4 是本发明一示例性实施例提供的一种基于心跳的数据同步装置的结构方框图；

图 5 是本发明一示例性实施例提供的一种基于心跳的数据同步装置的结构方框图；

图 6 是本发明一示例性实施例提供的一种基于心跳的数据同步装置的结构方框图。

具体实施方式

为使本发明的目的、技术方案和优点更加清楚，下面将结合附图对本发明实施方式作进一步地详细描述。

请参考图 1A，其示出了本发明一个示例性实施例提供的分布式存储系统的结构示意图。该分布式存储系统包括：若干个存储设备 110。

该分布式存储系统中存储有若干个数据块组，每个数据块组包含多块相同的数据块，且同一数据块组中的多个数据块存储于不同的存储设备中。在存储一个数据块组的多个存储设备中，有一个存储设备为存储该数据块组的主设备，其余的存储设备为存储该数据块组的从设备，主设备负责处理与客户端之间的交互。

在本发明实施例中，一个数据块组的主设备可以获取该数据块组的存取状态信息，根据该数据块组的存取状态信息确定该数据块组的心跳时间，根据该

数据块组的心跳时间向该从设备发送数据同步指令；该数据块组的从设备，用于根据该数据同步指令进行数据同步，以使得该多个数据块保持一致。

可选的，该分布式存储系统还包括：协调设备 120。协调设备 120 通过有线或者无线网络与若干个存储设备 110 相连并进行通信。

请参考图 1B，其示出了本发明示例性实施例涉及的一种网络设备的结构示意图。该网络设备 10 可以是上述存储设备 110 或者协调设备 120，该网络设备 10 包括：处理器 12 和网络接口 14。

处理器 12 包括一个或者一个以上处理核心。处理器 12 通过运行软件程序以及模块，从而执行各种功能应用以及数据处理。

网络接口 14 可以为多个，该网络接口 14 用于与其它存储设备或者网络设备进行通信。

可选的，网络设备 10 还包括存储器 16、总线 18 等部件。其中，存储器 16 与网络接口 14 分别通过总线 18 与处理器 12 相连。

存储器 16 可用于存储软件程序以及模块。具体的，存储器 16 可存储操作系统 162、至少一个功能所需的应用程序模块 164。操作系统 162 可以是实时操作系统（Real Time eXecutive，RTX）、LINUX、UNIX、WINDOWS 或 OS X 之类的操作系统。

请参考图 1C，其示出了本发明示例性实施例涉及的一种应用程序模块的示意图。如图 1C 所示，当网络设备 10 是一个存储设备，且该存储设备是一个数据块组的主设备时，该应用程序模块 164 可以是信息获取模块 164a、心跳时间确定模块 164b、第一指令发送模块 164c、第一时间发送模块 164d、第二时间发送模块 164e、第一时间接收模块 164f 和第二指令发送模块 164g。

信息获取模块 164a 获取数据块组的存取状态信息。

心跳时间确定模块 164b 根据数据块组的存取状态信息确定数据块组的心跳时间。

第一指令发送模块 164c 根据数据块组的心跳时间向从设备发送数据同步指令。

第一时间发送模块 164d 将数据块组的心跳时间发送给从设备。

第二时间发送模块 164e 将数据块组的心跳时间发送给协调设备。

第一时间接收模块 164f 接收协调设备返回的、修正后的心跳时间。

第二指令发送模块 164g 根据修正后的心跳时间向从设备发送数据同步指

令。

请参考图 1D，其示出了本发明示例性实施例涉及的另一种应用程序模块的示意图。当网络设备 10 是上述数据块组的从设备时，该应用程序模块 164 可以是指令接收模块 164h、数据同步模块 164i、第二时间接收模块 164j、设置模块 164k 和选举发起模块 164l。

指令接收模块 164h 接收主设备根据数据块组的心跳时间发送的数据同步指令；该心跳时间是由主设备获取数据块组的存取状态信息，并根据数据块组的存取状态信息确定。

数据同步模块 164i 根据数据同步指令进行数据同步，以使得多个数据块保持一致。

第二时间接收模块 164j 接收主设备发送的心跳时间。

设置模块 164k 根据心跳时间设置选举超时时间。

选举发起模块 164l 在从设备在选举超时时间内未接收到主设备发送的任何信号时，向该多个存储设备中的其它各个存储设备发起主设备选举。

请参考图 1E，其示出了本发明示例性实施例涉及的又一种应用程序模块的示意图。当网络设备 10 是上述协调设备 120 时，该应用程序模块 164 可以是统计模块 164m、第三时间接收模块 164n、等级确定模块 164p、修正模块 164q 和时间返回模块 164r。

统计模块 164m 统计数据块组的存取状态信息。

第三时间接收模块 164n 接收主设备发送的、数据块组的心跳时间；该心跳时间是由主设备获取数据块组的存取状态信息，并根据该数据块组的存取状态信息确定。

等级确定模块 164p 根据数据块组的存取状态信息确定该数据块组的重要性等级。

修正模块 164q 根据数据块组的重要性等级对心跳时间进行修正。

时间返回模块 164r 将修正后的心跳时间返回给主设备。

图 2 是本发明一示例性实施例提供的一种基于心跳的数据同步方法的方法流程图，该方法可以用于如图 1 所示的存储有至少一个数据块组的存储系统中。如图 2 所示，该基于心跳的数据同步方法可以包括：

步骤 201，主设备获取数据块组的存取状态信息。

其中，该数据块组的存取状态信息可以包括该数据块组的读频率和写频率。

该数据块组的读频率是存储系统中的主设备在一段时间内接受到对该数据块组中的数据块进行读取的操作的次数。该数据块组的写频率是存储系统中的主设备在一段时间内接受到对该数据块组中的数据块进行写入的操作的次数。

以基于 Raft 一致性算法的分布式存储系统为例，对于某一数据块组，客户端通常只对存储在该数据块组对应的主设备中的数据块进行读写操作，故主设备在一段时间内被读取或者写入数据的操作的次数在一定程度上，反映了一个数据组被使用的频繁程度。当一个存储设备成为一个数据块组的主设备之后，即可以统计该数据块组在一段时间内的读频率和写频率等存取状态信息。

可选的，该数据块组的存取状态信息还可以包括存储该数据块组的各个存储设备的工作状态信息，比如，该数据块组的存取状态信息还可以包括存储该数据块组的各个存储设备的数据吞吐量、CPU（Central Processing Unit，中央处理器）使用率、内存使用率和 I/O（Input/Output，输入/输出）占用率等。

在分布式存储系统中，一个存储设备的数据吞吐量、CPU 使用率、内存使用率和 I/O 占用率等信息可以反映该存储设备接收数据同步指令或者心跳信号的频繁程度。例如，一个存储设备的数据吞吐量大、CPU 使用率高、内存使用率高和 I/O 占用率高，则反映了该存储设备频繁接收数据同步指令或者心跳信号；反之，一个存储设备的数据吞吐量小、CPU 使用率低、内存使用率低和 I/O 占用率低，则反映了该存储设备接收数据同步指令或者心跳信号较少。

该步骤可以由图 1B 所示的网络设备 10 中的处理器 12 执行信息获取模块 164a 来实现。

步骤 202，主设备根据数据块组的存取状态信息确定该数据块组的心跳时间。

其中，一个数据块组对应的心跳时间越长，分布式存储系统中该数据块组的主设备和从设备之间单位时间通信次数越少；反之，心跳时间越短，该数据块组的主设备和从设备之间单位时间内通信次数越多。对于一个读写频率较低的数据块组，其主设备和从设备之间单位时间内进行数据同步的次数保持在一个较低的水平即可以满足该数据块组中的各个数据块的同步性能要求，因此，本发明实施例可以通过为各个数据块组分别设置独立的心跳时间，使读写次数较少的数据块组的主设备和从设备之间单位时间通信次数保持在一个较低的

水平，从而在保证数据块组的同步性能的同时，尽可能的减少数据块组的主设备和从设备之间单位时间内的通信次数，降低分布式存储系统的系统开销，提高存储系统的读写性能。

可选的，主设备根据数据块组的存取状态信息确定该数据块组的心跳时间的步骤，可以包括以下两种可能实现的方式。

1）主设备根据预设的第一评分规则存取状态信息进行评分，获得第一参考分值，根据该第一参考分值确定该数据块组的心跳时间。

在上述可能实现的方式中，以存取状态信息为数据块组的读频率和写频率为例，进行举例说明。首先，主设备根据预设的第一评分规则对数据块组的读频率和写频率进行评分，并将对读频率和写频率的评分的和作为第一参考分值，根据第一参考分值确定数据块组的心跳时间。

其中，第一评分规则可以是针对数据块组的存取状态信息预先指定的一个打分规则，例如，针对上述的读频率和写频率来说，打分规则可以是如下表 1 所示的规则。

表 1

第一评分规则	读/写频率	数据吞吐量	CPU 使用率	内存使用率	I/O 占用率
1 分	小于 50Hz	小于 800KB/s	小于 20%	小于 35%	小于 30%
2 分	50-100 Hz	800-2000KB/s	20%-80%	35%-70%	30%-70%
3 分	大于 100Hz	大于 2000KB/s	大于 80%	大于 70%	大于 70%

根据上述第一评分规则确定的数据块组的存取状态信息求和，作为第一参考分值。例如，数据块组的存取状态信息为读频率 70Hz 以及写频率 45Hz，若按照上述表格中表示的第一评分规则进行评分，第一参考分值为读频率 70Hz 对应的 2 分与写频率 45Hz 对应的 1 分之和，第一参考分值为 3 分。

其次，主设备根据第一参考分值 3 分确定数据块组的心跳时间。确定的方式可以根据第一参考分值的大小来确定，其中一种可能实现的确定方式可以参见表 2。

表 2

第一参考分值	心跳时间

1-3 分	5ms
4-7 分	2ms
7 分以上	1ms

主设备根据表 2 提供的第一参考分值与心跳时间的对应关系，找到第一参考分值 3 分对应的心跳时间 5ms，将 5ms 确定为数据块组的心跳时间。

需要特别说明的是，上述表 1 并不对第一评分规则形成限定，表 1 仅为第一评分规则的其中一种可能实现的方式。另外，表 1 中第二列中的存取状态信息“读/写频率”，表示读频率和写频率对应的评分规则相同，在实际应用中，读频率和写频率的评分规则也可以不同。

类似的，上述表 2 仅为第一参考分值与心跳时间的一种可能实现的对应方式，本实施例不限定第一参考分值与心跳时间的对应方式。

2）主设备根据数据块组的存取状态信息、参考时间以及第一权重计算数据块组的心跳时间。首先，主设备获取预先设置的参考时间以及第一权重；其次，主设备根据存取状态信息、该参考时间以及该第一权重计算该数据块组的心跳时间。

在上述可能实现的方式中，参考时间是一个通过机器学习获取的数值或者预先设定的数值。第一权重包含至少两个信息分别对应的权重。在计算心跳时间时，可以将各个存取状态信息与各自对应的权重相乘，或者，将各个存取状态信息进行预定处理后与各自对应的权重相乘，处理的方式可以是打分等方式，例如，通过上述第一评分规则进行处理，将得到的各个存取状态信息分别对应的乘积相加，得到一个和，之后，用预先设定好的参考时间除以上述各个存取状态信息分别对应的乘积相加得到的和，获得对应的心跳时间。

比如，以各个存取状态信息为读频率和写频率为例，在计算心跳时间时，可以根据下述公式计算心跳时间：

heartbeatTime= Time/(weightR*R+ weightW*W)；weightR+weightW=1。

在上述公式中，heartbeatTime 为心跳时间，Time 为参考时间，R 为读频率数值，W 为写频率数值，weightR 为读频率对应的权重，weightW 为写频率对应的权重。比如，以参考时间为 120ms，读频率为 70Hz，对应权重 0.2、写频率为 45Hz，对应权重 0.8 为例，根据上述计算公式计算获得的心跳时间为 120ms/(70*0.2+ 45*0.8)=2.4ms。

或者，设参考时间为 12ms，数据块组的存取状态信息包括以下五个数据：

读频率 70Hz、写频率 45Hz、数据吞吐量 900KB/s、CPU 使用率 40%以及内存使用率 50%；上述 5 个数据各自对应的权重为：0.05、0.05、0.1、0.2 和 0.4，根据上述计算方法获得心跳时间为：12ms/（70*0.05+45*0.05+0.9*01+0.4*0.2+0.5*0.4），约等于 2ms。

需要特别说明的是，本实施例仅以上述两种具体的实施场景为例对主设备根据数据块组的存取状态信息、参考时间以及第一权重计算数据块组的心跳时间的方式进行举例说明，并不对存取状态信息的类型和计算公式进行限定。

该步骤可以由图 1B 所示的网络设备 10 中的处理器 12 执行心跳时间确定模块 164b 来实现。

步骤 203，主设备将该数据块组的心跳时间发送给该从设备。

该步骤可以由图 1B 所示的网络设备 10 中的处理器 12 执行第一时间发送模块 164d 来实现。

步骤 204，从设备接收主设备发送的该心跳时间。

该步骤可以由图 1B 所示的网络设备 10 中的处理器 12 执行第二时间接收模块 164j 来实现。

步骤 205，从设备根据该心跳时间设置选举超时时间。

该步骤可以由图 1B 所示的网络设备 10 中的处理器 12 执行设置模块 164k 来实现。

步骤 206，从设备在选举超时时间内未接收到主设备发送的任何信号时，向该多个存储设备中的其它各个存储设备发起主设备选举。

以基于 Raft 一致性算法的分布式存储系统为例，系统根据选举的方式将存储设备划分为 Leader（领导者）、Follower（追随者）和 Candidate（候选者）三种角色，本方案中的主设备相当于 Leader、从设备相当于 Follower。在这三种角色中，Leader 负责日志的同步管理（即数据一致性管理）、处理来自客户端的包括读写操作在内的访问以及通过向 Follower 发送心跳保持联系；Follower 负责响应 Leader 的日志同步请求（即数据同步指令）、响应 Candidate 发起的投票请求以及将收到的客户端的请求转发给 Leader。

该选举超时时间用于触发从设备发起 Leader 选举，具体的，从设备中设置一定时器，从设备每次接受到主设备发送的心跳信号或者数据同步指令后，重置定时器，若该定时器计时到选举超时时间时仍未接收到心跳信号或者数据同步指令，则默认主设备发生故障，此时，从设备重新发起选举，发起选举的从

设备身份由 Follower 转变为 Candidate，并通过远程过程调用协议（Remote Procedure Call Protocol， RPC）向其它从设备发起投票请求（RequestVote），请求其它从设备支持该从设备成为 Leader，若该从设备接收到超过半数从设备的支持投票，则身份转为 Leader。

其中，从设备可以通过以下三种方式设置选举超时时间：

1）从设备根据预设的第二评分规则对该存取状态信息进行评分，获得第二参考分值，根据该第二参考分值确定该数据块组的第一超时系数，将该第一超时系数与该心跳时间的乘积设置为该选举超时时间。

其中，从设备根据预设的第二评分规则对该存取状态信息进行评分并获取第二参考分值的具体方法与主设备根据第一评分规则对该存取状态信息进行评分并获取第一参考分值的过程类似，此处不再赘述。

在本发明实施例中，从设备可以预先存储各个参考分值段与各个超时系数之间的对应关系，在根据该第二参考分值确定该数据块组的第一超时系数时，从设备可以查询该第二参考分值所在的参考分值段，并在预先存储的对应关系中查询与该参考分值段对应的第一超时系数。本发明实施例对于各个参考分值段与各个超时系数之间的对应关系的具体形式不进行限制，比如，请参考表 3，其示出了参考分值段与超时系数之间的对应关系表。

表 3

参考分值段	超时系数
1-2 分	70
3-4 分	60
5-6 分	50
7 分以上	40

如表 3 所示，假设从设备根据第二评分规则对该数据块组的存取状态信息进行评分并获取第二参考分值为 1.8，查询表 3 确定对应的第一超时系数为 70，主设备发送的心跳时间为 2ms，则可以设置选举超时时间为 70*2ms=140ms。

2）从设备获取预先设置的参考系数以及第二权重，该第二权重包含该存取状态信息分别对应的权重，根据该存取状态信息、该参考系数以及该第二权重计算该数据块组的第二超时系数，将该第二超时系数与该心跳时间的乘积设置为该选举超时时间。

在上述可能的实现方式中，参考系数可以是一个通过机器学习获取的数值，或者，也可以是一个预先设定的数值。第二权重包含数据块组的存取状态信息的权重。从设备可以将该存取状态信息对应的数值与该存取状态信息各自对应的权重分别相乘，将得到的乘积相加，得到一个和，该乘积相加得到的和再与参考系数相加即得到第二超时系数，第二超时系数与心跳时间的乘积即为选举超时时间。从设备根据该心跳时间设置选举超时时间的公式可以如下：

OverTime = （weightR*R+ weightW*W+Reference）* heartbeatTime;

weightR+weightW=1;

其中， OverTime 为选举超时时间， heartbeatTime 为心跳时间，Reference 为参考系数，R 为读频率的数值，W 为写频率的数值，weightR 为读频率对应的权重，weightW 为写频率对应的权重。

比如，在一种可能的实现方式中，数据块组的存取状态信息包括数据块组的读频率和写频率，且分别为 70Hz 和 45Hz，对应的权重分别为 0.2 和 0.8，参考系数为 15，心跳时间为 2ms，则电子设备根据上述方案可以计算该第二超时系数为 70*0.2+45*0.8+15=65，即选举超时时间为 65*2ms=130ms。

或者，在另一种可能的实现方式中，心跳时间为 2ms，参考系数为 15，数据块组的存取状态信息包括：读频率（70Hz）、写频率（45Hz）和 CPU 使用率（40%）；相对应的第二权重分别为：0.2、0.6 和 0.2，则电子设备根据上述方案可以计算该第二超时系数为 70*0.2+45*0.6+40*0.2+15=64，选举超时时间为 64*2ms=128ms。

3）从设备将预设的参考系数与该心跳时间的乘积设置为该选举超时时间。

从设备中也可以直接将参考系数与主设备发送的心跳时间相乘，将乘积作为选举超时时间。

类似于步骤 202，本发明实施例仅以上述具体的实施场景为例对从设备根据心跳时间和数据块组的存取状态信息计算选举超时时间的方式进行举例说明，并不对存取状态信息的类型和具体的计算公式构成限定。

该步骤可以由图 1B 所示的网络设备 10 中的处理器 12 执行选举发起模块 1641 来实现。

步骤 207，主设备根据该数据块组的心跳时间向该从设备发送数据同步指令。

该步骤可以由图 1B 所示的网络设备 10 中的处理器 12 执行第一指令发送

模块 164c 来实现。

步骤 208，从设备接收该主设备根据该数据块组的心跳时间发送的数据同步指令。

该步骤可以由图 1B 所示的网络设备 10 中的处理器 12 执行指令接收模块 164h 来实现。

步骤 209，从设备根据该数据同步指令进行数据同步。

在本发明实施例中，同一个数据块组中的各个数据块之间可以通过日志复制（Log Replication）来进行数据同步；在 Raft 一致性算法的分布式存储系统中，主设备当接收到客户端的日志（事务请求）后，先把该日志追加到本地的日志中，然后通过心跳将该日志同步给各个从设备，从设备接收并记录该日志后向主设备发送确认响应，当主设备收到一半以上的从设备返回的确认响应后，将该日志设置为已提交并追加到本地磁盘中并通知客户端，并在下个心跳通知从设备将该日志存储在自己的本地磁盘中。

该步骤可以由图 1B 所示的网络设备 10 中的处理器 12 执行数据同步模块 164i 来实现。

另外，本实施例提供的方法不限制使用在基于 Raft 一致性算法的分布式存储系统，也可以应用在其它基于心跳时间进行数据同步的分布式存储系统中。

综上所述，上述实施例提供的一种基于心跳的数据同步方法的方法，通过主设备获取数据块组的存取状态信息，根据数据块组的存取状态信息确定数据块组的心跳时间，根据数据块组的心跳时间向所从设备发送数据同步指令，指示从设备进行数据同步，解决了现有的 Raft 一致性算法中，一个数据块组所在的各个存储设备之间需要频繁的收发的信号，导致系统开销较大，影响存储系统的读写性能的问题，达到了降低分布式存储系统的系统开销，提高存储系统的读写性能的效果。

图 3 是本发明一示例性实施例提供的一种基于心跳的数据同步方法的方法流程图，该方法用于该方法可以用于如图 1 所示的包含协调设备，且存储有至少一个数据块组的存储系统中。如图 3 所示，该基于心跳的数据同步方法可以包括：

步骤 301 中，协调设备统计至少一个数据块组的存取状态信息。

其中，该至少一个数据块组的存取状态信息可以由该至少一个数据块组各

自的主设备发送给协调设备。

该步骤可以由图 1B 所示的网络设备 10 中的处理器 12 执行统计模块 164m 来实现。

步骤 302 中，协调设备根据该数据块组的存取状态信息确定该数据块组的重要性等级。

其中，协调设备可以根据数据块组存取状态信息中的读频率和写频率之和确定数据块组的重要性等级，比如，请参考表 4，其列出了某一存储系统中各个数据块组的读频率、写频率以及每个数据块组的重要性的排名。

表 4

数据块组名称	读频率	写频率	读写频率之和	重要性排名
第一数据块组	35Hz	25 Hz	60Hz	3
第二数据块组	20 Hz	15 Hz	35Hz	4
第三数据块组	40 Hz	50 Hz	90Hz	2
第四数据块组	55 Hz	45 Hz	100Hz	1

如表 4 所示，协调设备计算四个数据块组各自的读写频率之和，并按照四个数据块组各自的读写频率之和的大小确定四个数据块组的重要性，数据块组的读写频率之和越大，重要性越高，对应的重要性排名越靠前。

或者，协调设备也可以获取第三权重，该第三权重包括读频率和写频率各自对应的权重，协调设备根据各个数据块组对应的读频率、写频率以及该第三权重确定各个数据块组的重要性等级。

该步骤可以由图 1B 所示的网络设备 10 中的处理器 12 执行等级确定模块 164p 来实现。

步骤 303，一个数据块组的主设备向协调设备发送该数据块组的心跳时间。

该心跳时间是该主设备获取该数据块组的存取状态信息，并根据该数据块组的存取状态信息确定的心跳时间，该存取状态信息可以包括该数据块组的读频率和写频率在内的至少两个信息。主设备根据数据块组的存取状态信息确定

的心跳时间的步骤可以参考图 2 对应实施例中的描述，此处不再赘述。

该步骤可以由图 1B 所示的网络设备 10 中的处理器 12 执行第二时间发送模块 164e 来实现。

步骤 304，协调设备接收一个该主设备发送的、该数据块组的心跳时间。

该步骤可以由图 1B 所示的网络设备 10 中的处理器 12 执行第三时间接收模块 164n 来实现。

步骤 305 中，协调设备根据该数据块组的重要性等级对该数据块组的心跳时间进行修正。

在一种可能实现的方式中，修正心跳时间的方法可以如下：首先根据数据块组的重要性等级确定数据块组的修正系数，然后将修正系数与心跳时间的乘积确定为修正后的心跳时间。

例如，重要性从高到低的 5 个数据块组的修正系数分别为 0.7、0.8、1.0、1.2 和 1.5，心跳时间都为 2ms，则修正后的该 5 个数据块的心跳时间分别为 1.4ms、1.6ms、2ms、2.4ms 和 3.0ms。

该步骤可以由图 1B 所示的网络设备 10 中的处理器 12 执行修正模块 164q 来实现。

步骤 306 中，协调设备将修正后的心跳时间返回给该主设备，由该主设备根据该修正后的心跳时间向该从设备发送数据同步指令。

该步骤可以由图 1B 所示的网络设备 10 中的处理器 12 执行时间返回模块 164r 来实现。

在本发明实施例中，一个数据块组的主设备在生成心跳时间后，将该心跳时间发送给协调设备进行修正，在接收到协调设备返回的修正后的心跳时间之前，该主设备根据生成的心跳时间进行数据同步指令或者心跳信号的发送，在接收到协调设备返回的修正后的心跳时间之后，该主设备改为根据修正后的心跳时间进行数据同步指令或者心跳信号的发送。进一步的，主设备还将修正后的心跳时间发送给该数据块组的从设备，使该数据块组的从设备根据该修正后的心跳时间重新设置选举超时时间。

综上所述，本发明实施例提供的一种基于心跳的数据同步方法，通过协调设备统计数据块组的存取状态信息，根据数据块组的读存取状态信息确定数据块组的重要性等级，接收一个数据块组的主设备发送的、该数据块组的心跳时间，根据该数据块组的重要性等级对该数据块组的心跳时间进行修正，将修正

后的心跳时间返回给主设备，根据各个数据块组各自运行参数对各个数据块组的心跳时间进行修正，从整体上对存储系统各个数据块组的心跳时间进行优化，进一步提高存储系统的读写性能。

下述为本发明装置实施例，可以用于执行本发明方法实施例。对于本发明装置实施例中未披露的细节，请参照本发明方法实施例。

图 4 是本发明实施例提供的一种基于心跳的数据同步装置的结构方框图，该基于心跳的数据同步装置可以通过软件、硬件或者两者的结合实现成为分布式存储系统的存储设备中的部分或者全部。该分布式存储系统可以是 1A 所示的分布式存储系统，该存储设备是一个数据块组的主设备，该数据块组包含多个数据块，该多个数据块分别存储于该分布式存储系统的多个存储设备中，该多个存储设备包含该主设备，且该多个存储设备中除该主设备之外的其余设备为该数据块组的从设备，该基于心跳的数据同步装置可以包括：信息获取单元 401、心跳时间确定单元 402、第一指令发送单元 403、第一时间发送单元 404 和第二时间发送单元 405、第一时间接收单元 406 和第二指令发送单元 407。

信息获取单元 401，具有与信息获取模块 164a 相同或相似的功能。

心跳时间确定单元 402，具有与心跳时间确定模块 164b 相同或相似的功能。

第一指令发送单元 403，具有与第一指令发送模块 164c 相同或相似的功能。

第一时间发送单元 404，具有与第一时间发送模块 164d 相同或相似的功能。

第二时间发送单元 405，具有与第二时间发送模块 164e 相同或相似的功能。

第一时间接收单元 406，具有与第一时间接收模块 164f 相同或相似的功能。

第二指令发送单元 407，具有与第二指令发送模块 164g 相同或相似的功能。

图 5 是本发明实施例提供的一种基于心跳的数据同步装置的结构方框图，该基于心跳的数据同步装置可以通过软件、硬件或者两者的结合实现成为分布

式存储系统的存储设备中的部分或者全部。该分布式存储系统可以是 1A 所示的分布式存储系统，该存储设备是一个数据块组的从设备，该数据块组包含多个数据块，该多个数据块分别存储于该分布式存储系统的多个存储设备中，该多个存储设备包含该主设备，且该多个存储设备中除该主设备之外的其余设备为该数据块组的从设备，该基于心跳的数据同步装置可以包括：指令接收单元 501、数据同步单元 502、第二时间接收单元 503、设置单元 504 和选举发起单元 505。

指令接收单元 501，具有与指令接收模块 164h 相同或相似的功能。

数据同步单元 502，具有与数据同步模块 164i 相同或相似的功能。

第二时间接收单元 503，具有与第二时间接收模块 164j 相同或相似的功能。

设置单元 504，具有与设置模块 164k 相同或相似的功能。

选举发起单元 505，具有与选举发起模块 164l 相同或相似的功能。

图 6 是本发明实施例提供的一种基于心跳的数据同步装置的结构方框图，该基于心跳的数据同步装置可以通过软件、硬件或者两者的结合实现成为分布式存储系统的协调设备中的部分或者全部。该分布式存储系统可以是 1A 所示的分布式存储系统，该分布式存储系统存储有至少一个数据块组，一个数据块组包含多个数据块，该多个数据块分别存储于该分布式存储系统的多个存储设备中，该多个存储设备包含该主设备，且该多个存储设备中除该主设备之外的其余设备为该数据块组的从设备，该基于心跳的数据同步装置可以包括：统计单元 601、第三时间接收单元 602、等级确定单元 603、修正单元 604 和时间返回单元 605。

统计单元 601，具有与统计模块 164m 相同或相似的功能。

第三时间接收单元 602，具有与第三时间接收模块 164n 相同或相似的功能。

等级确定单元 603，具有与等级确定模块 164p 相同或相似的功能。

修正单元 604，具有与修正模块 164q 相同或相似的功能。

时间返回单元 605，具有与时间返回模块 164r 相同或相似的功能。

应当理解的是，在本文中使用的，除非上下文清楚地支持例外情况，单数形式“一个”（“a”、“an”、“the”）旨在也包括复数形式。还应当理解的是，在本

文中使用的“和/或”是指包括一个或者一个以上相关联地列出的项目的任意和所有可能组合。

上述本发明实施例序号仅仅为了描述，不代表实施例的优劣。

本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成，也可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，上述提到的存储介质可以是只读存储器，磁盘或光盘等。

以上所述仅为本发明的较佳实施例，并不用以限制本发明，凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

权 利 要 求 书

1、一种分布式存储系统，其特征在于，所述分布式存储系统存储有至少一个数据块组，且所述分布式存储系统包括多个存储设备；所述多个存储设备中的一个设备为存储所述数据块组的主设备，其余设备为存储所述数据块组的从设备；

所述主设备，用于获取所述数据块组的存取状态信息，根据所述数据块组的存取状态信息确定所述数据块组的心跳时间，根据所述数据块组的心跳时间向所述从设备发送数据同步指令；

所述从设备，用于根据所述数据同步指令进行数据同步。

2、根据权利要求 1 所述的系统，其特征在于，所述数据块组的存取状态信息包括所述数据块组的读频率和写频率。

3、一种基于心跳的数据同步装置，其特征在于，应用于分布式存储系统，所述分布式存储系统存储有至少一个数据块组，且所述分布式存储系统包括多个存储设备；所述多个存储设备中的一个设备为存储所述数据块组的主设备，其余设备为存储所述数据块组的从设备，所述主设备包括所述装置，所述装置包括：

信息获取单元，用于获取所述数据块组的存取状态信息；

心跳时间确定单元，用于根据所述数据块组的存取状态信息确定所述数据块组的心跳时间；

第一指令发送单元，用于根据所述数据块组的心跳时间向所述从设备发送数据同步指令，所述数据同步指令用于指示所述从设备进行数据同步。

4、根据权利要求 3 所述的装置，其特征在于，所述数据块组的存取状态信息包括所述数据块组的读频率和写频率。

5、根据权利要求 3 或 4 所述的装置，其特征在于，所述心跳时间确定单元，具体用于根据预设的第一评分规则对所述存取状态信息进行评分，获得第一参考分值，根据所述第一参考分值确定所述数据块组的心跳时间。

6、根据权利要求 4 所述的装置，其特征在于，所述心跳时间确定单元，具

体用于根据所述第一评分规则对所述读频率和所述写频率分别进行评分，并将对所述读频率和所述写频率的评分的和作为第一参考分值，根据所述第一参考分值确定所述数据块组的心跳时间。

7、根据权利要求 4 所述的装置，其特征在于，所述心跳时间确定单元，具体用于获取预先设置的参考时间以及所述存取状态信息对应的第一权重，根据所述存取状态信息、所述参考时间以及所述第一权重计算所述数据块组的心跳时间。

8、根据权利要求 7 所述的装置，其特征在于，所述心跳时间确定单元，具体用于通过以下公式计算所述心跳时间：

heartbeatTime= Time/(weightR*R+ weightW*W)； weightR+weightW=1；

其中，heartbeatTime 为所述心跳时间，Time 为所述参考时间，R 为所述读频率的数值，W 为所述写频率的数值，weightR 为所述读频率对应的权重，weightW 为所述写频率对应的权重。

9、根据权利要求 3-8 任一项所述的装置，其特征在于，所述装置还包括：

第一时间发送单元，用于将所述数据块组的心跳时间发送给所述从设备，使得所述从设备根据所述心跳时间设置选举超时时间。

10、根据权利要求 3-9 任一项所述的装置，其特征在于，所述分布式存储系统中还包含与所述多个存储设备相连接的协调设备，所述装置还包括：

第二时间发送单元，用于将所述数据块组的心跳时间发送给所述协调设备；

第一时间接收单元，用于接收所述协调设备返回的、修正后的心跳时间；

第二指令发送单元，用于根据所述修正后的心跳时间向所述从设备发送数据同步指令。

11、一种基于心跳的数据同步装置，其特征在于，应用于分布式存储系统，所述分布式存储系统存储有至少一个数据块组，且所述分布式存储系统包括多个存储设备；所述多个存储设备中的一个设备为存储所述数据块组的主设备，其余设备为存储所述数据块组的从设备，所述从设备包括所述装置，所述装置包括：

指令接收单元，用于接收所述主设备根据所述数据块组的心跳时间发送的数据同步指令，其中，所述主设备获取所述数据块组的存取状态信息，并根据所述数据块组的存取状态信息确定所述心跳时间；

数据同步单元，用于根据所述数据同步指令进行数据同步。

12、根据权利要求 11 所述的装置，其特征在于，所述数据块组的存取状态信息包括所述数据块组的读频率和写频率。

13、根据权利要求 12 所述的装置，其特征在于，所述装置还包括：

第二时间接收单元，用于接收所述主设备发送的所述心跳时间；

设置单元，用于根据所述心跳时间设置选举超时时间；

选举发起单元，用于当所述从设备在所述选举超时时间内未接收到所述主设备发送的任何信号时，向所述多个存储设备中的其它各个存储设备发起主设备选举。

14、根据权利要求 13 所述的装置，其特征在于，所述设置单元，具体用于根据预设的第二评分规则对所述存取状态信息进行评分，获得第二参考分值，根据所述第二参考分值确定所述数据块组的第一超时系数，将所述第一超时系数与所述心跳时间的乘积设置为所述选举超时时间。

15、根据权利要求 13 所述的装置，其特征在于，所述设置单元，具体用于获取预先设置的参考系数以及所述存取状态信息对应的第二权重，根据所述存取状态信息、所述参考系数以及所述第二权重计算所述数据块组的第二超时系数，将所述第二超时系数与所述心跳时间的乘积设置为所述选举超时时间。

16、根据权利要求 15 所述的装置，其特征在于，所述设置单元，具体用于通过以下公式设置所述选举超时时间：

OverTime = （weightR*R+ weightW*W+Reference）* heartbeatTime；

weightR+weightW=1；

其中， OverTime 为所述选举超时时间， heartbeatTime 为所述心跳时间，Reference 为所述参考系数，R 为所述读频率的数值，W 为所述写频率的数值，

weightR 为所述读频率对应的权重，weightW 为所述写频率对应的权重。

17、一种基于心跳的数据同步装置，其特征在于，应用于分布式存储系统，所述分布式存储系统存储有至少一个数据块组，且所述分布式存储系统包括多个存储设备以及与所述多个存储设备相连接的协调设备，所述多个存储设备中的一个设备为存储所述数据块组的主设备，其余设备为存储所述数据块组的从设备，所述协调设备包括所述装置，所述装置包括：

统计单元，用于统计所述数据块组的存取状态信息；

第三时间接收单元，用于接收所述主设备发送的、所述数据块组的心跳时间，其中，所述主设备获取所述数据块组的存取状态信息，并根据所述数据块组的存取状态信息确定所述心跳时间；

等级确定单元，用于根据所述数据块组的存取状态信息确定所述数据块组的重要性等级；

修正单元，用于根据所述数据块组的重要性等级对所述心跳时间进行修正；

时间返回单元，用于将修正后的心跳时间返回给所述主设备。

18、根据权利要求 17 所述的装置，其特征在于，所述数据块组的存取状态信息包括所述数据块组的读频率和写频率。

19、一种基于心跳的数据同步方法，其特征在于，用于分布式存储系统中，所述分布式存储系统存储有至少一个数据块组，且所述分布式存储系统包括多个存储设备；所述多个存储设备中的一个设备为存储所述数据块组的主设备，其余设备为存储所述数据块组的从设备，所述方法包括：

所述主设备获取所述数据块组的存取状态信息；

所述主设备根据所述数据块组的存取状态信息确定所述数据块组的心跳时间；

所述主设备根据所述数据块组的心跳时间向所述从设备发送数据同步指令，所述数据同步指令用于指示所述从设备进行数据同步。

20、根据权利要求 19 所述的方法，其特征在于，所述数据块组的存取状态信息包括所述数据块组的读频率和写频率。

21、根据权利要求 19 或 20 所述的方法，其特征在于，所述主设备根据所述数据块组的存取状态信息确定所述数据块组的心跳时间，包括：

所述主设备根据预设的第一评分规则对存取状态信息进行评分，获得第一参考分值；

所述主设备根据所述第一参考分值确定所述数据块组的心跳时间。

22、根据权利要求 20 所述的方法，其特征在于，所述主设备根据所述数据块组的存取状态信息确定所述数据块组的心跳时间，包括：

所述主设备根据所述第一评分规则对所述读频率和所述写频率分别进行评分；

所述主设备将对所述读频率和所述写频率的评分的和作为第一参考分值；

所述主设备根据所述第一参考分值确定所述数据块组的心跳时间。

23、根据权利要求 20 所述的方法，其特征在于，所述主设备根据所述数据块组的相关信息确定所述数据块组的心跳时间，包括：

所述主设备获取预先设置的参考时间以及所述存取状态信息对应的第一权重；

所述主设备根据所述存取状态信息、所述参考时间以及所述第一权重计算所述数据块组的心跳时间。

24、根据权利要求 23 所述的方法，其特征在于，所述主设备根据所述数据块组的存取状态信息确定所述数据块组的心跳时间，包括：

通过以下公式计算所述心跳时间：

heartbeatTime= Time/(weightR*R+ weightW*W)； weightR+weightW=1；

其中，heartbeatTime 为所述心跳时间，Time 为所述参考时间，R 为所述读频率的数值，W 为所述写频率的数值，weightR 为所述读频率对应的权重，weightW 为所述写频率对应的权重。

25、根据权利要求 19-24 任一项所述的方法，其特征在于，所述方法还包括：

所述主设备将所述数据块组的心跳时间发送给所述从设备，使得所述从设

备根据所述心跳时间设置选举超时时间。

26、根据权利要求 19-25 任一项所述的方法，其特征在于，所述分布式存储系统还包含与所述多个存储设备相连接的协调设备，所述方法还包括：

所述主设备将所述数据块组的心跳时间发送给所述协调设备；

所述主设备接收所述协调设备返回的、修正后的心跳时间；

所述主设备根据所述修正后的心跳时间向所述从设备发送数据同步指令。

27、一种基于心跳的数据同步方法，其特征在于，用于分布式存储系统中，所述分布式存储系统存储有至少一个数据块组，且所述分布式存储系统包括多个存储设备；所述多个存储设备中的一个设备为存储所述数据块组的主设备，其余设备为存储所述数据块组的从设备，所述方法包括：

所述从设备接收所述主设备根据所述数据块组的心跳时间发送的数据同步指令，其中，所述主设备获取所述数据块组的存取状态信息，并根据所述数据块组的存取状态信息确定所述心跳时间；

所述从设备根据所述数据同步指令进行数据同步。

28、根据权利要求 27 所述的方法，其特征在于，所述数据块组的存取状态信息包括所述数据块组的读频率和写频率。

29、根据权利要求 28 所述的方法，其特征在于，所述方法还包括：

所述从设备接收所述主设备发送的所述心跳时间；

所述从设备根据所述心跳时间设置选举超时时间；

当所述从设备在所述选举超时时间内未接收到所述主设备发送的任何信号时，所述从设备向所述多个存储设备中的其它各个存储设备发起主设备选举。

30、根据权利要求 29 所述的方法，其特征在于，所述从设备根据所述心跳时间设置选举超时时间，包括：

所述从设备根据预设的第二评分规则对所述存取状态信息进行评分，获得第二参考分值；

所述从设备根据所述第二参考分值确定所述数据块组的第一超时系数；

所述从设备将所述第一超时系数与所述心跳时间的乘积设置为所述选举超时时间。

31、根据权利要求 29 所述的方法，其特征在于，所述从设备根据所述心跳时间设置选举超时时间，包括：

所述从设备获取预先设置的参考系数以及所述存取状态信息对应的第二权重；

所述从设备根据所述存取状态信息、所述参考系数以及所述第二权重计算所述数据块组的第二超时系数；

所述从设备将所述第二超时系数与所述心跳时间的乘积设置为所述选举超时时间。

32、根据权利要求 31 所述的方法，其特征在于，所述从设备根据所述心跳时间设置选举超时时间，包括通过以下公式设置所述选举超时时间：

OverTime = （weightR*R+ weightW*W+Reference）* heartbeatTime；

weightR+weightW=1；

其中， OverTime 为所述选举超时时间， heartbeatTime 为所述心跳时间，Reference 为所述参考系数，R 为所述读频率的数值，W 为所述写频率的数值，weightR 为所述读频率对应的权重，weightW 为所述写频率对应的权重。

33、一种基于心跳的数据同步方法，其特征在于，用于分布式存储系统中，所述分布式存储系统存储有至少一个数据块组，且所述分布式存储系统包括多个存储设备以及与所述多个存储设备相连接的协调设备；所述多个存储设备中的一个设备为存储所述数据块组的主设备，其余设备为存储所述数据块组的从设备，所述协调设备执行所述方法，所述方法包括：

所述协调设备统计所述数据块组的存取状态信息，并接收所述主设备发送的、所述数据块组的心跳时间，其中，所述主设备获取所述数据块组的存取状态信息，并根据所述数据块组的存取状态信息确定所述心跳时间；

所述协调设备根据所述数据块组的存取状态信息确定所述数据块组的重要性等级；

所述协调设备根据所述数据块组的重要性等级对所述心跳时间进行修正；

所述协调设备将修正后的心跳时间返回给所述主设备。

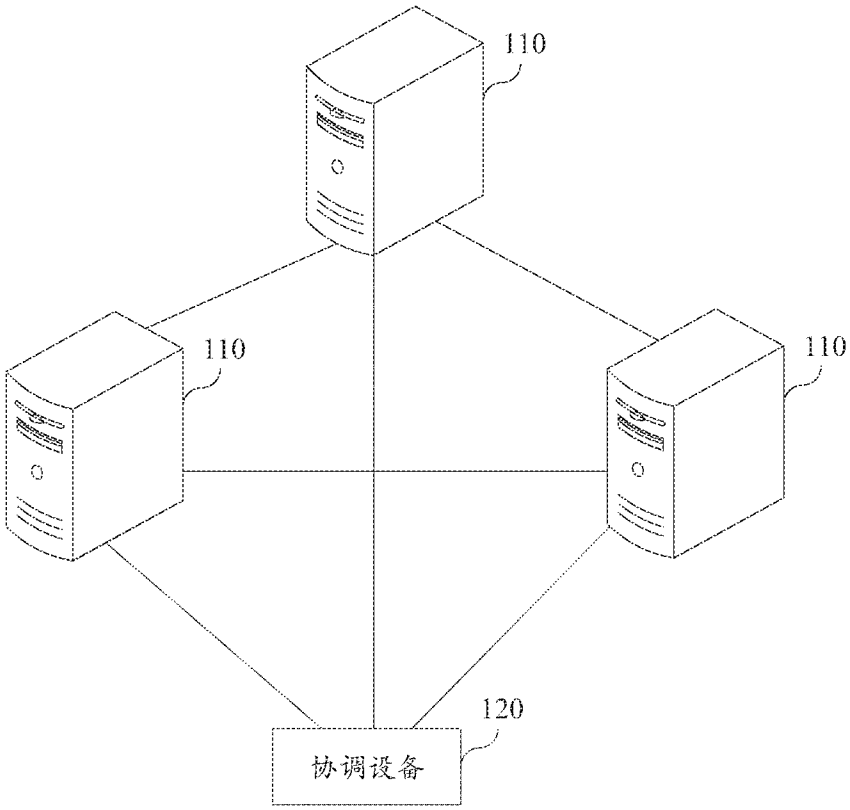

110
110
110
120
协调设备

图 1A

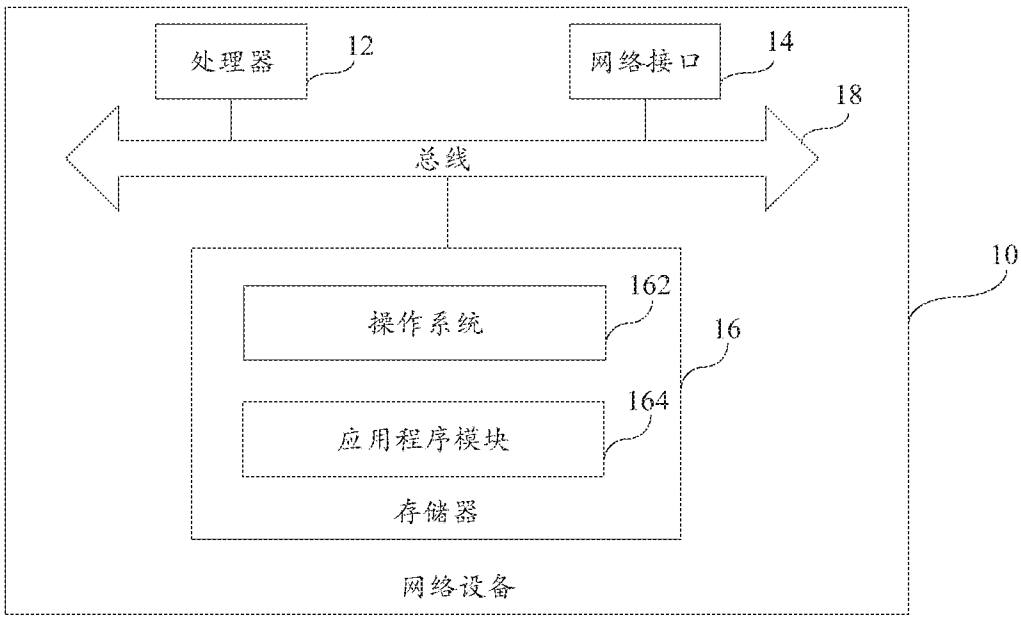

处理器
12
网络接口
14
18
总线
10
162
操作系统
16
164
应用程序模块
存储器
网络设备

图 1B

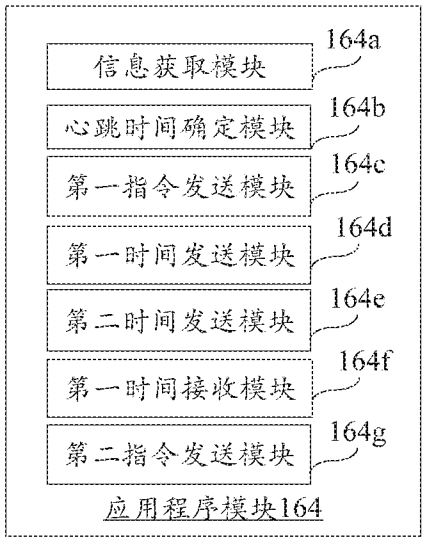

信息获取模块
164a
心跳时间确定模块
164b
第一指令发送模块
164c
第一时间发送模块
164d
第二时间发送模块
164e
第一时间接收模块
164f
第二指令发送模块
164g
应用程序模块164

图 1C

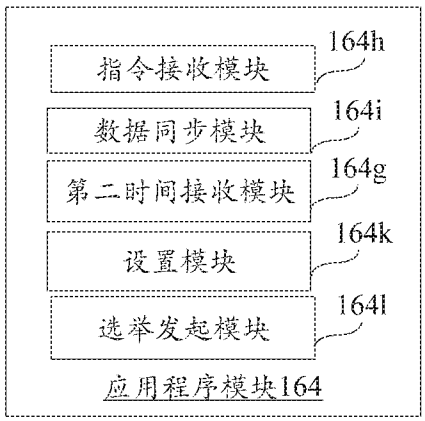

指令接收模块
164h
数据同步模块
164i
第二时间接收模块
164g
设置模块
164k
选举发起模块
164l
应用程序模块164

图 1D

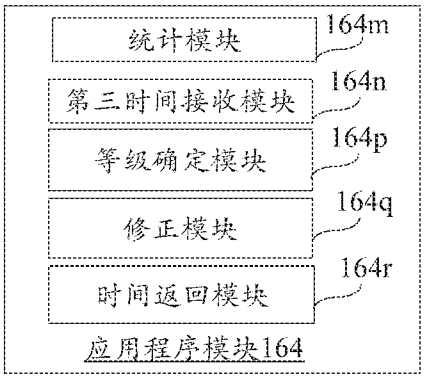

统计模块
164m
第三时间接收模块
164n
等级确定模块
164p
修正模块
164q
时间返回模块
164r
应用程序模块164

图 1E

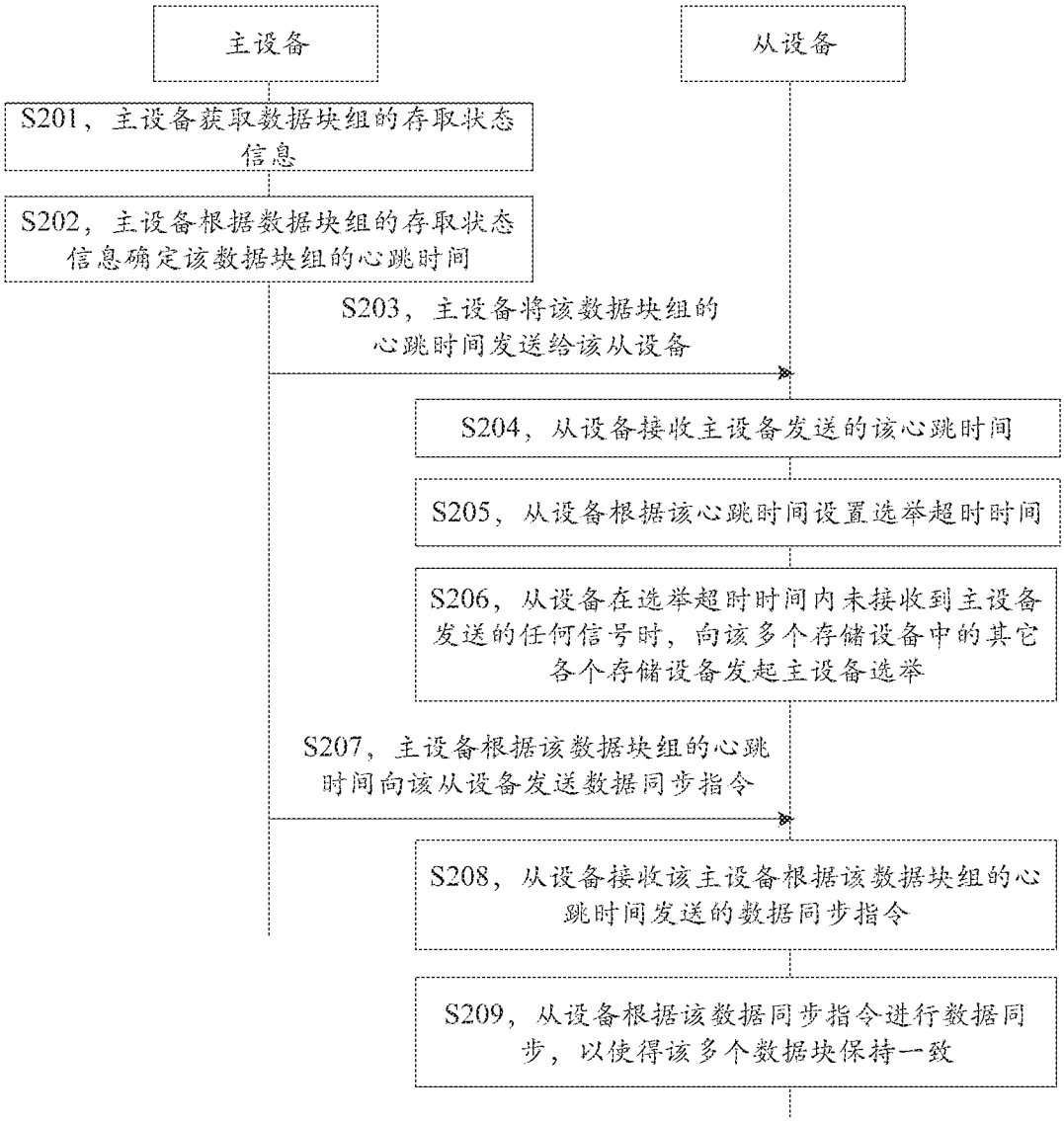

主设备
从设备
S201，主设备获取数据块组的存取状态信息
S202，主设备根据数据块组的存取状态信息确定该数据块组的心跳时间
S203，主设备将该数据块组的心跳时间发送给该从设备
S204，从设备接收主设备发送的该心跳时间
S205，从设备根据该心跳时间设置选举超时时间
S206，从设备在选举超时时间内未接收到主设备发送的任何信号时，向该多个存储设备中的其它各个存储设备发起主设备选举
S207，主设备根据该数据块组的心跳时间向该从设备发送数据同步指令
S208，从设备接收该主设备根据该数据块组的心跳时间发送的数据同步指令
S209，从设备根据该数据同步指令进行数据同步，以使得该多个数据块保持一致

图 2

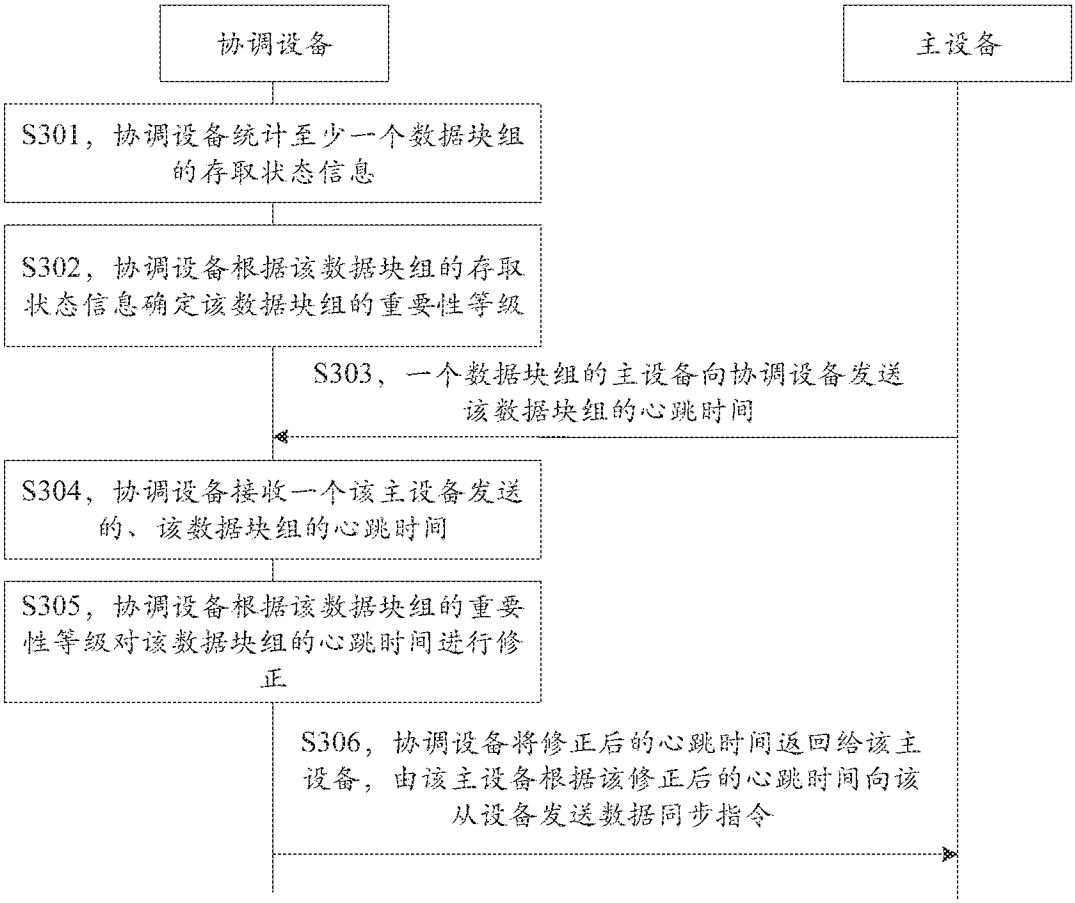

协调设备
主设备
S301，协调设备统计至少一个数据块组的存取状态信息
S302，协调设备根据该数据块组的存取状态信息确定该数据块组的重要性等级
S303，一个数据块组的主设备向协调设备发送该数据块组的心跳时间
S304，协调设备接收一个该主设备发送的、该数据块组的心跳时间
S305，协调设备根据该数据块组的重要性等级对该数据块组的心跳时间进行修正
S306，协调设备将修正后的心跳时间返回给该主设备，由该主设备根据该修正后的心跳时间向该从设备发送数据同步指令

图 3

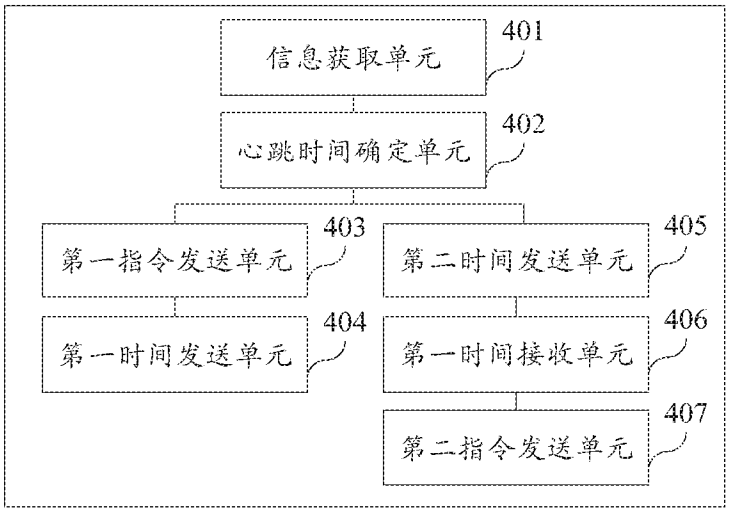

信息获取单元
401
心跳时间确定单元
402
第一指令发送单元
403
第二时间发送单元
405
第一时间发送单元
404
第一时间接收单元
406
第二指令发送单元
407

图 4

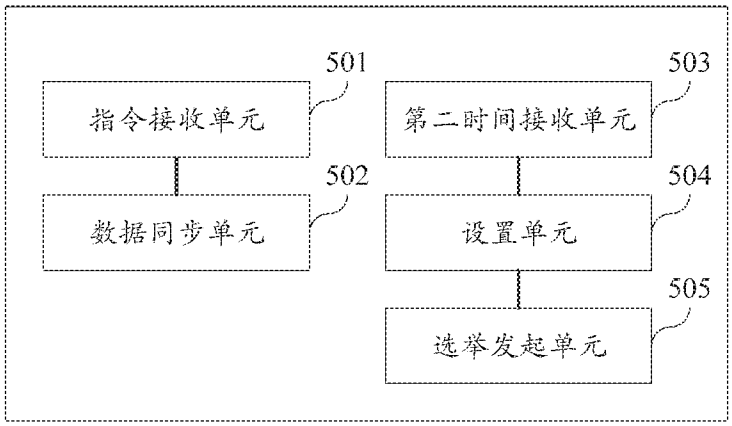

501
指令接收单元
502
数据同步单元
503
第二时间接收单元
504
设置单元
505
选举发起单元

图 5

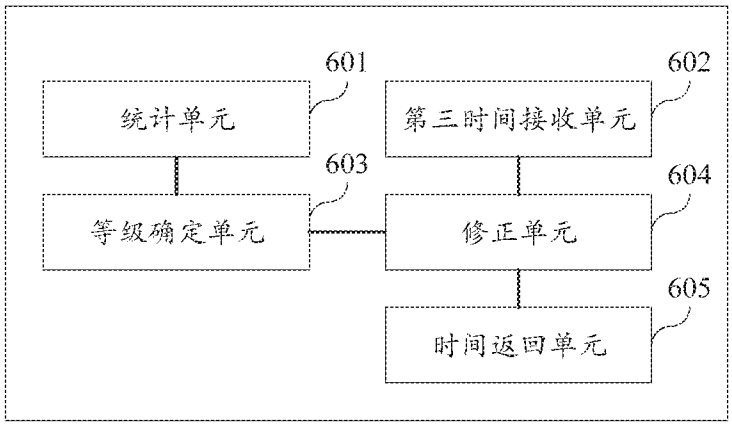

601
统计单元
602
第三时间接收单元
603
等级确定单元
604
修正单元
605
时间返回单元

图 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/097244

A. CLASSIFICATION OF SUBJECT MATTER

H04L 29/08 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

H04L, G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, WPI, EPODOC, CNKI: discrete, distribut+, cloud+, stor+, heartbeat?, time, period, master, primary, main, slav+, other, synchroniz+, updat+, replicat+, copy+, read+, writ+, access+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 104994168 A (SUZHOU KEDA TECHNOLOGY CO., LTD.), 21 October 2015 (21.10.2015), description, paragraphs [0008]-[0013]	1-33
A	CN 104765661 A (SHENZHEN ANYUN INFORMATION TECHNOLOGY CO., LTD. et al.), 08 July 2015 (08.07.2015), the whole document	1-33
A	US 2015074178 A1 (SAMSUNG ELECTRONICS CO., LTD.), 12 March 2015 (12.03.2015), the whole document	1-33

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
01 November 2016 (01.11.2016)	**14 November 2016 (14.11.2016)**
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer **CHENG, Jiali** Telephone No.: (86-10) **62413289**

Form PCT/ISA/210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016/097244

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 104994168 A	21 October 2015	None	
CN 104765661 A	08 July 2015	None	
US 2015074178 A1	12 March 2015	KR 20150030036 A	19 March 2015

国际检索报告

国际申请号
PCT/CN2016/097244

A. 主题的分类

H04L 29/08(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

H04L, G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNPAT, WPI, EPODOC, CNKI: 分布式, 分立式, 云, 存储, 心跳, 时间, 周期, 主, 从, 从属, 同步, 复制, 读, 写, 存取, distribut+, cloud+, stor+, heartbeat?, time, period, master, primary, main, slav+, other, synchroniz+, updat+, replicat+, copy+, read+, writ+, access+

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
A	CN 104994168 A (苏州科达科技股份有限公司) 2015年 10月 21日 (2015-10-21) 说明书第[0008]-[0013]段	1-33
A	CN 104765661 A (深圳市安云信息科技有限公司等) 2015年 7月 8日 (2015-07-08) 全文	1-33
A	US 2015074178 A1 (SAMSUNG ELECTRONICS CO., LTD.) 2015年 3月 12日 (2015-03-12) 全文	1-33

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2016年 11月 1日	2016年 11月 14日

ISA/CN的名称和邮寄地址	受权官员
中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088	程佳丽
传真号 (86-10)62019451	电话号码 (86-10)62413289

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2016/097244

检索报告引用的专利文件	公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN 104994168 A	2015年 10月 21日	无	
CN 104765661 A	2015年 7月 8日	无	
US 2015074178 A1	2015年 3月 12日	KR 20150030036 A	2015年 3月 19日